SMU Data Science Review

Volume 5 | Number 2 Article 3

2021

Identification and Characterization of Forest Fire Risk Zones Leveraging Machine Learning Methods

Joshua Balson
Southern Methodist University, jbalson8@gmail.com

Matt Chinchilla
Southern Methodist University, Mattchinchilla@gmail.com

Cam Lu
Southern Methodist University, camlu2010@gmail.com

Jeff Washburn
Southern Methodist University, jwashbur65@gmail.com

Nibhrat Lohia
Southern Methodist University, nlohia@smu.edu

Identification and Characterization of Forest Fire Risk Zones Leveraging Machine Learning Methods

Joshua Balson [1], Matthew Chinchilla [1], Cam Lu [1], Jeff Washburn [1], Nibhrat Lohia [2]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] Adjunct Lecturer at Southern Methodist University,
Dallas, TX 75275 USA

Abstract. Across the United States, record numbers of wildfires are observed costing billions of dollars in property damage, polluting the environment, and putting lives at risk. The ability of emergency management professionals, city planners, and private entities such as insurance companies to determine if an area is at higher risk of a fire breaking out has never been greater. This paper proposes a novel methodology for identifying and characterizing zones with increased risks of forest fires. Methods involving machine learning techniques use the widely available and recorded data, thus making it possible to implement the tool quickly.

1 Introduction

Fires happen naturally; this study is not meant to prevent all fires but limit destruction to private property, loss of life and understand the air quality implications. Fires have an impact on property, life, and health aspects regarding air quality. People are encroaching on burned areas—keeping areas free from public and private buildings where they do not belong. Belcher et al. (2014) contend that smoldering combustion leads to mega-fires as measured in the whole natural fabric devoured [6]. The creators review the current information on smoldering fires within the soil framework for combustion flow and chemistry, highlighting contrasts with blazing fires. Belcher et al. (2014) suggest that smoldering combustion of characteristic ground powers leads to the biggest and longest burning fires, like peatlands.

California has faced a record-breaking number of wildfires in the last decade, and this trend will continue. Seven of the ten most destructive wildfires in California history occurred in the previous five years that consumed over 1 million acres of land, destroyed 32 thousand structures, and killed 128 people. In 2020, 7,606 wildfires burned 2,277,922 acres of land and damaged or destroyed 3,849 structures. Unfortunately, the severity of wildfires in California is worsening. In addition to the personal and economic toll resulting from destructive wildfire activity, Wildfire smoke can impact the health of communities and people near and far. Most healthy people will recover quickly from wildfire smoke exposure and will not suffer long-term health consequences. However, infants and children, people with pre-existing medical conditions such as lung and heart conditions, and socioeconomic factors to access health care may be more sensitive to wildfire smoke. Wildfires leave behind toxic

debris in the air, soil, and waterways and have a long-lasting costly impact on human health.

California Wildfires' total economic cost was $85 billion in 2017, $400 billion in 2018, and $80 billion in 2019. The changing climate has been a critical factor in increasing the risk and extent of wildfires in California. Wildfire risk depends on several factors, including global warming, temperature, soil moisture, wind, and the presence of trees, shrubs, and other potential fuels. All these factors have substantial direct and indirect ties to climate variability and climate change. One-degree Celsius temperature increase would raise the median burned area per year by as much as 600 percent in some forest types. Once a fire starts (people cause more than 80 percent of U.S. wildfires), warmer temperatures, high wind, and drier conditions can help fires spread and make them harder to put out. In the first week of President Biden's administration, he announced a $2 trillion climate plan to address the climate crisis. He put climate change at the center of domestic, national security, and foreign policy. His actions reflect the urgency of fighting climate change.

Wildfires cause massive destruction, fatalities, displacement of people from their homes and cities. In addition to these, the effects may include deteriorating air conditions risking the safety of inhabitants. This is particularly evident in the American Southwest. Beginning in the late nineteen eighties, the Western part of the United States has experienced a prolonged drought period. Since that time, large fires have occurred at two-year intervals and have caused significant financial losses and other sociological impacts to the inhabitants of this region.

This study aims to identify high-risk areas of wildfires using distinct aspects of relating factors, including but not limited to economic and climate data, leveraging machine learning methods. The research will also identify the key indicators in predicting wildfires and the related effects as part of the process. This study's results may assist communities, builders, homeowners, and forest managers in decision-making and policy design.

Wildfires are a developing common risk in most United States regions, posing a threat to life and property, especially where local biological systems meet developed areas. Nevertheless, since the fire could be a normal (and regularly advantageous) process, fire suppression can lead to more extreme fires due to the buildup of vegetation, which makes more fuel. Also, the secondary impacts of wildfires, including disintegration, avalanches, the introduction of intrusive species, and water quality changes, are frequently more disastrous than the fire itself.

Three components are significant for creating a rapidly spreading fire: fuel, oxygen, and heat. Firefighters regularly allude to this combination of elements as the fire triangle. On a sweltering day, drought conditions peak, and something as little as a spark has the potential to make a huge, rapidly spreading fire with devastating results. VanSomeren et al. (2020) stated that although fire can be caused by the sun or a lightning strike, most wildfires are a result of human carelessness. Unextinguished campfires, lit cigarette butts, improperly burned debris, and arson is responsible for 84% of wildfires [26]. Man-made, rapidly spreading fires have tripled the fire season from 46 days to 154 days with a staggering cost of $2 billion. Once considered a natural phenomenon started by lighting, wildfires are presently being recognized due to human mistakes.

The federal government yearly spends billions of dollars to stifle wildfires. Wildfires increase the potential for flooding, debris flows, and landslides. Smoke and other emissions contain toxins that can cause critical health issues. Short-term impacts: pulverization of timber, forage, wildlife habitats, picturesque vistas, and watersheds. Long-term impacts: reduced access to recreational regions; destruction of community infrastructure and cultural and financial assets

Wildfires can have immediate and long-term impacts on the quality of streams, lakes, and other bodies of water. The foremost recognizable effect of wildfires is stormwater runoff. After the loss of vegetation, the ground's soil repels water and avoids the retention of water. This failure to retain water advances debris transportation and silt into bigger bodies of water, further contaminating valuable and essential resources. Post-fire flash floods become a danger and permit heavy metals from cinder and soil to penetrate waterways. Sifting these water sources can be expensive as well as time-consuming. Depending on the temperature and time of year a wildfire occurs, vegetation can be impacted. Plants on the timberland floor or smaller trees are frequently crushed by wildfires, whereas bigger trees can outlive if the fire does not spread into the tree canopy. The flames from these fires devastate the nourishment source and homes of numerous creatures, debilitating their survival. Plant and trees that can survive the flares are vulnerable to illness, organisms, and insects due to their diminished resistance after burn wounds. Wildfires have both immediate and long-term impacts on air quality. As a forest burns, expansive sums of smoke are discharged into the environment. These smoke particles are regularly small and made up of gasses and water vapor. Air contamination from fires can travel extraordinary distances and often may posture a risk to human health. These tiny particles can end up held up deep inside the lungs, making it troublesome to breathe and placing extra stress on the heart.

Moreover, wildfires deliver an expanded sum of carbon monoxide, which can also lead to an assortment of health implications. Structures that lie inside the way of a wildfire are crushed, uncovering dangerous materials that pose a danger to human health for first responders and during the cleanup handle. More seasoned homes built before the 1970s regularly contain a mineral called asbestos. Once asbestos is aggravated, the strands become airborne and, when breathed in, can lead to the development of pleural mesothelioma within the lining of the lungs. Amid the cleanup process, numerous materials are often improperly disposed of and threaten devastation in the future. These cleanup processes must be dealt with safely, completed with the right gear, and done by professionals.

Although wildfires leave an immense sum of devastation in their path, they do leave behind a few valuable qualities as well. Numerous plants require routine burns to spread seeds and survive. Fires can kill illnesses and insects, which will improve the livelihood of plants by removing excess debris from the timberland floor and permitting more access to the supplements given by uncovered daylight. Low concentrated fires clear underbrush and help deter future fires from spreading. After a wildfire, new meadows are made and permit brushing animals to benefit from the change. This increment within the standard arrangement of species permits a shift in biology that promotes growth and the nonstop cycle of life. Vegetation, just like the fireweed, requires the disturbance of fire to sprout and allow the regrowth of plants that have died due to the fire. As plants and vegetation die away, life starts to mend and spring forth.

While numerous wildfires cause negligible harm to the land and pose few dangers to the land or individuals downstream, some fires cause harm that requires extraordinary efforts to avoid issues afterward. Loss of vegetation exposes soil to erosion; water runoff may increase and cause flooding; silt may move downstream and harm houses or fill stores putting endangered species and community water supplies at risk. After a fire, the primary need is crisis stabilization to avoid further harm to life, property, or natural resources. The stabilization work starts promptly and may proceed for up to a year. The longer-term recovery effort to repair the harm caused by the fire begins after the fire is out and proceeds for several years. Recovery focuses on the lands unlikely to recoup naturally from wildfires.

Unlike numerous characteristic fiascos, most fierce blazes can be avoided. The natural and prudent costs of fierce blazes have an effect that endures for a long time. Through utilizing caution, taking preventative measures, and observing fires mindfully, the dangers related to these devastating tragedies can be lowered.

Fires cause widespread destruction with catastrophic aftermath of the loss of life, billions of dollars in property damage, and health implications due to environmental pollution. The ability to characterize and classify areas at higher risk levels of forest fires can limit the loss of life, property damage and improve quality of life by allowing emergency responders to have a proactive response rather than a reactive response to wildfires.

2 Literature Review

Previous work on predicting forest fires using meteorological data was performed in Europe. Cortez et al. (2007) used SVM and four meteorological inputs to predict the burned area of small fires [10]. Forest fires are a major environmental issue, creating economic and ecological damage while endangering human lives. Fast detection is a crucial element for controlling fires. An alternative is to use automatic tools based on local sensors, such as meteorological stations. Meteorological conditions like wind, temperature, humidity, and elevation are known to influence forest fires. Several fire indexes, such as the Forest Fire Weather Index (FWI), use such data. This study leverages machine learning methods to identify and characterize forest fire risk levels for a specific geographic area. This study focused on various machine learning methods using a feature set that includes meteorological input to predict an area's potential fire risk. Such knowledge is beneficial for improving firefighting resource management for prioritizing high-risk regions to clear the land or for fire and ground crews to concentrate on a particular area.

According to Andrews et al. (2007), one significant concern is forest fires, which influence woodland conservation, cause substantial environmental harm, and cause human anguish [8]. Human carelessness or natural causes such as lightning lead to forest fires. Despite an increment of state costs to control fires, a great many forest fires are annihilated every year. Specifically, California is profoundly influenced by forest fires. In 2015, many sections of land were destroyed, and some fatalities occurred.

According to Cortez et al. (2007), fast detection is a crucial element for successful firefighting. Since traditional human surveillance is expensive and affected by

subjective factors, there has been an emphasis on developing automated solutions, including satellite-based, infrared/smoke scanners, and local sensors. Satellites have acquisition costs, localization delays, and the resolution is not adequate for all cases. Moreover, scanners have high equipment and maintenance costs. Weather conditions, such as temperature and air humidity, affect fire occurrence. Since automatic meteorological stations are often available, such data can be collected in real-time, with low costs.

Cortez et al. (2007) mentioned that meteorological data had been incorporated into numerical indices used for prevention and support fire management decisions. Such decisions include the level of readiness, prioritizing targets, or evaluating guidelines for safe firefighting. In particular, the Canadian Forest Fire Weather Index (FWI) system was designed in the 1970s when computers were scarce; thus, it required only simple calculations using look-up tables with readings from four meteorological observations (i.e., temperature, relative humidity, rain, and wind) that could be manually collected in weather stations. Nevertheless, this index is universally used in Canada and several countries around the world.

One additional candidate explanatory variable in the initial model will be a measure of vegetation or shrub growth that has accumulated over a given area. It has long been assumed that the age of shrub growth is correlated with higher risks of wildfires. However, Moritz et al. (2004) found that the relative period of shrub growth does not seem to be a significant factor in fire frequencies in the American West [15]. This is especially the case in Southern and Central California, where the Santa Ana winds, which occur each fall and blow from east to west, appear to create extreme fire dangers solely based on the strength, elevated temperature, and low humidity of these wind currents. This suggests that vegetation and shrub growth may not strongly predict fire risk, especially in areas where other factors can be powerful enough to override other potential leading indicators for high-risk fires completely.

Despite the findings of Moritz et al., there is other evidence to suggest that the relationship between the time since the fire and vegetation age can be an important contributing factor to fire risk [15]. Fauria and Johnson (2008) point out that "many studies have noted changes in the dominant pattern of fire variability and fire-climate relationships from centennial/decadal to interannual time scales" (pg. 2325). Therefore, it may be possible that the relationships between climate, vegetation, and fire risk change over time based on changes in weather patterns that occur at even intervals – and therefore could be meaningful as an input variable to a fire risk model.

Since the question of how much (or how little) vegetation age can affect fire risk is still an undeveloped area of research, it seems prudent to include vegetation data in the preliminary stages of building the fire risk model. Ghorbanzde et al. (2019) noted that forest ecosystems contain about 66% of the total terrestrial carbon. They are a significant component of the global carbon budget. Forests are a vital component to maintaining an environmental balance and regulating the carbon cycle [1]. If vegetation is not a significant indicator, it can simply be removed at a later stage of refining the model. Either way, it will help lead to further insights towards understanding the relationships between vegetation age and fire.

Forest wildfires, caused naturally or by humans, are the most dangerous and destructive disasters. They are hard to predict and hard to extinguish. Many search efforts have been conducted to monitor, predict, and prevent wildfires using Artificial

Intelligence techniques. Sayad et al. (2019), in their research article *Predictive modeling of wildfires: A new dataset and machine learning approach*, elaborate their research efforts to prevent wildfires by predicting the occurrence of wildfires using Big Data, Machine Learning, and Remote Sensing, which related to Normalized Differential Surface Index (NDVI) of the quantify vegetation greenness, Land Surface Temperature (LST) of the meteorological condition and fire indicator of thermal anomalies [4]. The data were acquired from the Moderate Resolution Imaging Spectroradiometer, an instrument on the Terra and Aqua satellites. High prediction accuracy of 98.32% was achieved using validation strategies such as classification metrics, cross-validation. The combination of Artificial Intelligence technologies (Big Data, Machine Learning, and Data Mining) and Internet of Things technologies (wireless sensors and cameras) help improve wildfire monitoring and prediction systems. Toulouse et al. (2009) took a different approach using computer vision to build a fire pixel detection model with infrared and visible images [3]. The dataset consists of 500 images of wildfire collected worldwide and 100 multi-modal images. The prediction model relies on a publicly available database containing many images of wildfire with various dominated fire colors, conditions of smoke, environments, and background. Toulouse et al. (2009) has shown a method promising for detecting fire in real-world situations [3].

One form of climate data collected for this study is drought data from the National Integrated Drought Information System (NIDIS), a part of the National Oceanic and Atmospheric Administration (NOAA). Drought is highly correlated with forest and wildfire in Sen, Z's (2015) book Applied drought modeling, prediction, and mitigation drought is defined as the temporary reduction in the rainfall, runoff, and soil moisture amounts in an area. In the literature, Sen points out that the effects of drought are more harmful in humid regions than arid regions because settlers in humid areas are not prepared for the impacts of water scarcity. Drought has economic and environmental effects, including reduced crop yield, increase in plant diseases, deterioration in water quality, soil erosion, forest fires, and wildfires. An area that begins to experience precipitation lower than the long-term average for an extended amount of time is considered drought area [7].

Different indexes are available to measure the severity of drought, but the most popular is the Palmer Drought Severity Index (PDSI) and the Standardized Precipitation Index (SPI). The PDSI was the first comprehensive and most frequently applied meteorological multivariable drought index in the United States. It is based on the measurement of moisture supply departures from normal conditions, which is calculated from the water balance of a two-layer soil model using monthly mean precipitation and temperature data as well as the locally available water content of the soil [7]. The SPI is based on rainfall alone and relies on a long-term precipitation record, generally at least a 30-year normal duration. The long-term record is fitted to a probability density function such as the Gamma or logarithmic normal, so the fitted distribution corresponds to the same percentile on a normal distribution. The SPI is applicable because its standardization represents the same probability of drought occurrence, regardless of the time, location, and climate [7]. The PDSI and SPI drought indicators are both included as features of the NIDIS data collected.

How have the climate changes impacted forest fires? Flannigan et al. (2000) states six components influencing forest fires: fire frequency, size, intensity, seasonality, type,

and severity [22]. Fire frequency affects ecosystems through interrupting or terminating individuals' life cycles such as soil structure, carbon storage and nutrient cycling. Fire size determines landscape patchiness and distance seed will have to travel for regeneration. Fire intensity is the amount of energy released during a fire burn. It depends on the accumulation rate of fuel it consumes and loading, topography, meteorological influences, and characteristics of the previous disturbance [22]. Time of year may affect fire intensity through differences in surface fuel moisture contents. The seasonal phenological state of forest plants burned will determine the vegetative reproductive response and influence post-fire ecosystems. Fire-type can vary across the area of burn that might be initiated by ground fires, crown fires, or surface fires [22]. Ground fires occur below materials such as leaves; they are slow-moving but can take out large areas, crown fires pose a considerable risk due to fast spread behavior that can jump from one tree to another, and surface fires are the tamest fire and create the least amount of destruction. In addition to climate influences on the fire regime, other factors such as ignition agent, fire season length, vegetation characteristics, and human activities such as fire management policies and landscape fragmentation may influence the fire regime in the next century [22]. The impact of climate change on the fire regime could significantly impact ecosystems due to increases in area burned and fire severity. Manipulations of fuel type, load, and arrangement could be used to help protect areas. But at a large scale, fuel management would not be feasible [22]. A delicate balancing act is required by land management to protect values from fire while at the same time allowing the fire to resume its natural role in ecosystem functioning.

The Witch Creek Fire was the second-largest wildfire of the 2007 California wildfire season and the largest of October's 2007 California wildfires. Burning 197,990 acres (about the area of Austin, Texas) alone, after merging with the Poomacha and McCoy fires, the Witch-Guejito-Poomacha Complex Fire had a total burn area of 247,800 acres. Witch Creek fire led to evacuations of 500,000 people (about half the population of South Dakota), becoming the largest evacuation in California history. Michael et al. 2010, state when the Witch Creek Fire began ravaging San Diego County, California, in October 2007, there was no way of knowing how far it would spread, how many homes it would destroy, how many families would be evacuated, or how many livers would be lost [21]. People were not going to wake up and realize that it was a bad dream. This "Witch" was real and was devouring everything in her path. No matter how prepared you may believe you are for an evacuation under emergency circumstances. You are not [21]. Cedar Fire is an opportunity to test evacuation plans prepared for a full-scale evacuation, but people quickly learned that they were wrong. Wildfire is a part of nature. It plays a key role in shaping ecosystems by serving as an agent of renewal and change. Fire can be deadly, destroying homes and polluting the air with emissions harmful to human health. Therefore, keeping people well-informed about the risks of wildfire and safely and effectively evacuate an area is essential to reduce risk. A California brush fire can spread the distance of a football field in less than 4 seconds [21]. If the fires force inhabitants to evacuate, it would be necessary to wear masks or wrap bandanas around the mouths and noses to filter out the smoke and ash. However, people are not well preparing when the fires force them to leave. There are not nearly enough crates to transport dogs, cats, puppies, and kittens [21]. Not enough Personal Protective Equipment supplies such as masks. Neighboring firefighters try to give 2 hours' notice, but realistically people would have minutes to evacuate. Is there enough

warning to get out? What would a person need to take? How much could they take with them? How long would it take to load? How much could they take? Would the fire be nipping at their tails as people dashed down to the road? The answers to these questions were based on gut instinct and the broadcasts of the fire's progress [21]. As the smoke and floating ash drew nearer, time was of the essence. Fire, flooding, earthquakes, and toxic threats must be considered in developing an evacuation plan. But based on past experiences, no matter how much is planned, it is never enough.

Cheuklap et al. (2018) study vegetation regeneration after wildfires between Mediterranean (Witch Creek Fire) and tundra (Anaktuvuk River Fire) ecosystems [23]. After eight years of disturbance, the vegetation in the Mediterranean ecosystem had still not yet recovered, whereas the tundra ecosystem retrieved in just three years. Higher degree burning leads to a quicker vegetation recovery rate. However, ecological retrogression was detected. Spatial heterogeneity in post-fire vegetation recovery was observed in both sites. Wildfires have devastating effects on the environment, ecology, and economy [23]. A loss of vegetation reduces biodiversity, promotes the invasion of exotic species, enhances soil erosion, and even increases the potential for flooding and debris flows [23]. Witch Creek Fire in California and Anaktuvuk reiver Fire in Alaska have these similarities and differences in post-fire vegetation regeneration. They had similar shrubs and grasses, and the fires had similar burn times. Most of the vegetation within the burn scar is 84% shrub/scrub, with forest areas of just 14% and 2% grasslands. A study showed that significant vegetation recovery in the year following the wildfires [23]. However, eight years after the Witch Creek Fire, the area's vegetation regeneration index had still not returned to its original level due to the implication of ecological structures that did not resume after the wildfire [23]. Compared with the Witch Creek Fire, the Anaktuvuk River Fire area showed a more dynamic change, and the vegetation regeneration index reached pre-fire condition after the fire. The regeneration index of the Anaktuvuk River Fire area in terms of burn severity was opposite that in Witch Creek Fire. The highest degree of burn severity had the lowest regeneration index. A low degree of burn severity had a higher regeneration index because the vegetation was not entirely consumed by the wildfire [23].

3 Methods

The primary goal of the algorithm being trained in this study is to predict areas at higher risk of forest and/or wildfire. Climate data will be used primarily to train the algorithm to predict the fire risk areas. Geo-mapping overlay data of urban density and building footprints have also been collected to enhance the usefulness of these predictions. In association with NASA's Biological Diversity and Ecological Forecasting program, Microsoft's Bing map team has compiled a U.S.-wide vector building data set with over 125 million building footprints in the United States. The data set includes total footprint coverage and the number of unique buildings intersecting each cell within a 30m raster layer cell. The study also collected population density data from a survey conducted by the American Economic Association. The urban population data calculate two measures of density the native density of population per square kilometer and experienced density population within ten

kilometers of the average resident. The urban population data covers U.S. metropolitan areas for the entire United States. As population centers and cities expand into uninhabited areas, people may unintentionally build or settle in places experiencing a high fire risk. Adding the additional overlay of build footprint and urban density can help highlight property and populations that may need to consider proactive measures to protect themselves from forest and wildfires should they find themselves in areas of increased risk.

The collection of data for the study posed considerable challenges not initially expected. Climate data collected by government agencies and private entities is generally collected to create maps and images. While much of this data is distributed freely, many file formats are not easily reverted from a geo-mapping format to a format conducive to machine learning. An example of this challenge is data produced by Landfire, a shared program between the wildland fire management programs of the U.S. Department of Agriculture Forest Service and the U.S. Department of the Interior, providing landscape-scale geospatial products to support cross-boundary planning, management, and operations. Landfire produces surface fuel and vegetation layered map data that is freely available. Still, the data is distributed in a format exclusive to Landfire (.lcp) and is only accessible using FlamMap, a fire analysis desktop application. Another unexpected challenge in the collection of data for the study was stale climate data. Climate data is collected over a long period. Over time, providers of this data may update technology and change methodologies.

In some cases, the data may only be collected over a set time period. Many data sources and providers were reviewed in the search for consistent, quality, climate data. For this study, consistent and quality data is defined as data that has been collected since at least the year 2000, and quality is defined as the data is taken at regular intervals such as daily, weekly, or monthly and not missing large amounts of data over those periods.

Despite these challenges, the National Oceanic and Atmospheric Administration (NOAA) contains many different sources of quality time-series data for the locations desired, most of which are provided in CSV and text file format. Since the focus of our study will be on the Cleveland National Forest in San Diego County, data were downloaded from NOAA's websites for precipitation [https://www.ncdc.noaa.gov/cdo-web/search] and drought indices [https://www.drought.gov]. Data was also obtained for humidity, air pressure, and cloud cover [https://www.ncdc.noaa.gov/cdo-web/datatools/lcd]. The precipitation data was gathered from a weather station in Julian, CA, a small city that lies within Cleveland National Forest, at an elevation of 1,284 feet. Historical drought data was collected for San Diego County since this data is only provided at the county level. The remaining climate data (humidity, air pressure, cloud cover) was collected from a weather station at Miramar Naval Air Station, which lies 32.7 miles to the southwest of Cleveland National Forest. All these data sources span a period from January 1st, 2000, through March 6th, 2021, and contain no notable data gaps within this time. Ideally, all the data would be taken from different instrument locations within the park boundaries. However, this data was not easily obtainable from web sources that are available free to the public.

Before training the model, the area of interest is separated into a grid using the Military Grid Reference System (MGRS). MGRS can create a grid zone on a mapped area and is easily converted to latitude and longitude. Figure 2 below are examples of MGRS in use. Using the MGRS, evenly spaced grid points in southern California are

plotted. The overall area covered by the MGRS grid will consist of the ten counties San Luis Obispo, Kern, San Bernadino, Santa Barbara, Ventura, Los Angeles, Orange, Riverside, San Diego, and Imperial. (Fig. 1) The points plotted will become the local areas that the model will be trained to observe. For each grid point in the area of interest, the model will need to predict a high probability of a fire occurring.

Fig. 1. Area of focus for this study – Southern California

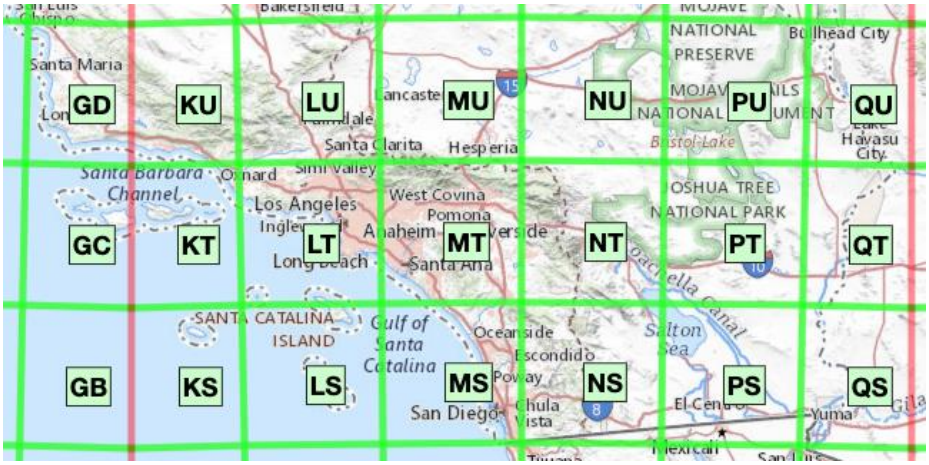

Fig. 2. MGRS coordinates of Southern California

Training data consisting of historical weather data, including atmospheric information such as temperature, precipitation, wind, and other conditions, was combined with a Kaggle data set that includes historic locations where fires have occurred make up time series-based training data. The predictive variable is a boolean value of zero or one. Zero indicates no fire in the given area for a given date, and one suggests that a fire had occurred.

Missing data are imputed using historical seasonal trends for a given area. For example, given a missing temperature value for a given date, the median temperature

for that missing date will be filled in with the historical average for the same date or period. From there, the data is split between a training set and a test set. The test set consists of a random sample of twenty-five percent of the entire initial data set. The training and test set data are then scaled separately to a unit standard deviation and run through multiple classifiers using five-fold cross-validation, each with multiple hyper-parameter settings to find the optimal settings for each classifier. Classifiers will be gauged for effectiveness on the test set by three metrics: accuracy, precision, and recall. Accuracy measures all predictions the classifier got right divided by all the observations in the model. Precision is a measure of the ratio of false positives to total predicted, and recall is a measure of the ratio of false negatives to the actuals. Below is an outline of the various classification, regression, and neural network methods that will be explored using the methods above.

The first and most critical component in creating the classification model was identifying the inputs and expected outputs. The main objective of the model used in this study is the binary classification if a fire will happen a certain number of days out. In this study, five lag days are used to predict if a fire will happen; therefore, five classification models are created for each lag day. The lag represents the number of days before if a fire is predicted to occur. For example, the lag-1 model predicts if a fire will happen tomorrow, while the lag-5 model predicts if a fire will happen five days from now for a specific geographic region. The model's input data consist of past weather data for each geographic location, the past 3 years of fires in each of those locations for when past fires have occurred, and the population density for the Southern California area. As shown in figure 3, the number of fires and the size of those fires encountered in Southern California from 2000 to 2015 are significant.

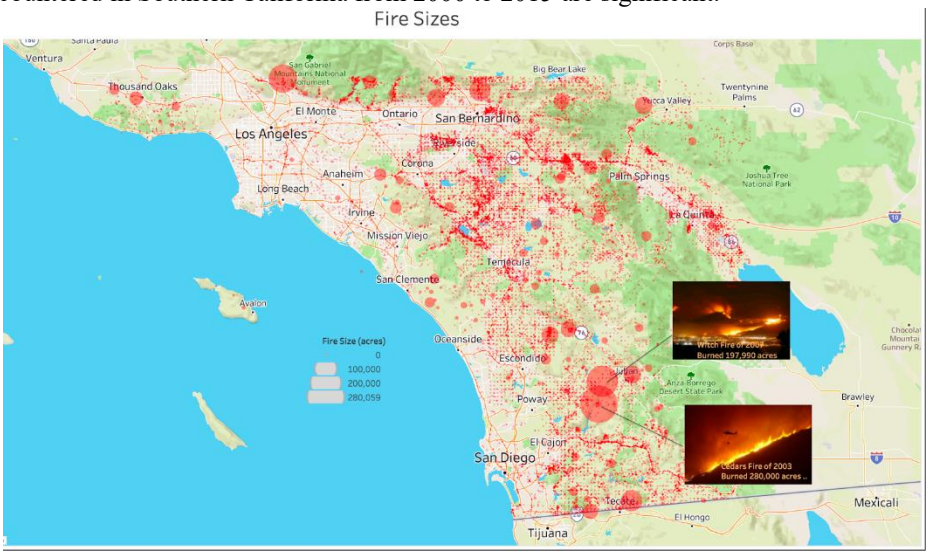

Fig. 3. All forest fires in the San Diego region between 1992 and 2015.

Using this past data, binary classification models were created to predict future fires in geographic locations. The expected improvement is to provide government and fire officials a longer lead time on the probability of a fire occurring in a specific area to take proactive measures to prevent the fire by clearing the debris and brush, or evacuating people to save lives. The measurement this study found most compelling in predicting future fires was achieving the highest F1 score. Precision and recall were considered and are good indicators of model performance; however, the imbalance of the training dataset in which fires were seen over the time period for each of the geographic locations required a focus on both precision and recall; hence the F1 score was used for evaluating model performance.

The data used for this study involved a Kaggle dataset which included 1.88 million georeferenced fires collected over 24 years. In addition to the fire data, the corresponding historical weather data for each geographic area were also collected. Lastly, that data was combined with the population density for the same areas. With the vast amount of fire data, this study focused on a subset of that data for Southern California using the historic fires from 2000 to 2015. Further sub-setting of the fire data included the filtering of fires that were less than a quarter of an acre and the winter months when fires were not likely. To create the geographic regions, each fire latitude and longitude coordinates were converted to a Military Grid Reference System (MGRS). MGRS solved the problem of defining a geographic location for a specific area and was used to correlate multiple fires to a specific area and ascertain the weather for that same area using an application programming interface (API) call to Meteomatics. As shown below in figure 4, the MGRS grid system helped define a geographic area.

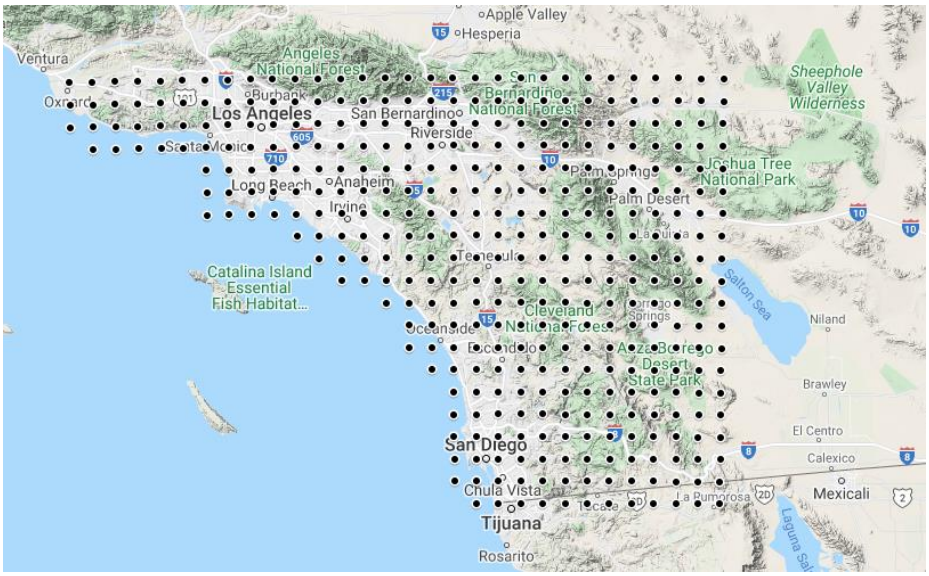

Fig. 4. MGRS coordinate grid for Southern California.

Taking this grid system and converting the past fires' latitude and longitude to the MGRS grid allowed for an easy way to associate the weather for that specific area when the fires occurred, as shown below in figure 5.

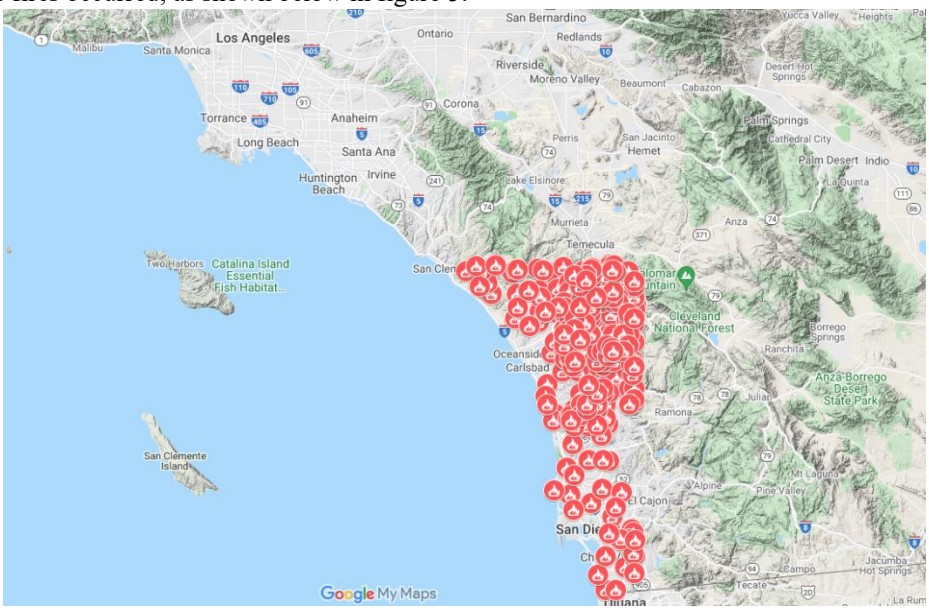

Fig. 5. Overlay of a section of fires to the MGRS grid system

MGRS proved extremely useful and solved the problem of defining an area for which a prediction is made. This is important because the focus of this study is predicting the probability of a fire. The predictions can be more specific to a smaller geographic area by representing smaller regions, providing greater precision on the area potentially affected by a fire.

Since the dataset was heavily imbalanced with most days for each geographic area not having fires, the model performance and success measure was based on the F1 score. The F1 score is the harmonic mean of precision and recall. A portion of the dataset was held out for testing purposes. The hold-out dataset consisted of the last three years of fire data from 2013 to 2015.

The remaining dataset was used for training that consisted of fires and historic weather from 2000 to 2012. One of the challenges to overcome was the significant imbalance of the training data containing much fewer days that had fires versus the days that did not. Therefore, an under-sampling of the training data was employed to train on a dataset containing a balance of days that had fires and those that did not have fires. Some feature engineering was needed to calculate the fire lag for the previous five days. While the original dataset had an is_fire feature, the predictions were required for future days; therefore, five new features were created - is_fire_lag1, is_fire_lag2, is_fire_lag3, is_fire_lag4, and is_fire_lag5, which represent the target feature for prediction of each of the models. In addition, the aggregation of fires for the past three years for the specific geographic area was created to help identify areas

that had higher fire rates to give a more weighted aspect to that specific area. The correlation matrix seen in figure 6 shows that the new features do not correlate with the other features.

Fig. 6. Correlation matrix

The model was trained using five-fold cross-validation on the undersampled data using the historical weather data and the engineered features.

Shown in figure 7 below are the average F1 Score metrics for each candidate classifier on a training data set that was withheld before training the model, which contained only the date ranges between the years of 2013 through 2015. The bars in red indicate higher recall scores, and the bars in blue indicate lower recall scores. Thicker bars represent higher precision scores, while thin bars represent lower precision scores. The logistic regression classifier was chosen because of its higher recall scores and its ability to return predicted probabilities. The higher recall was preferred over precision due to the desire to minimize prediction errors where fires occur in areas not predicted by the model.

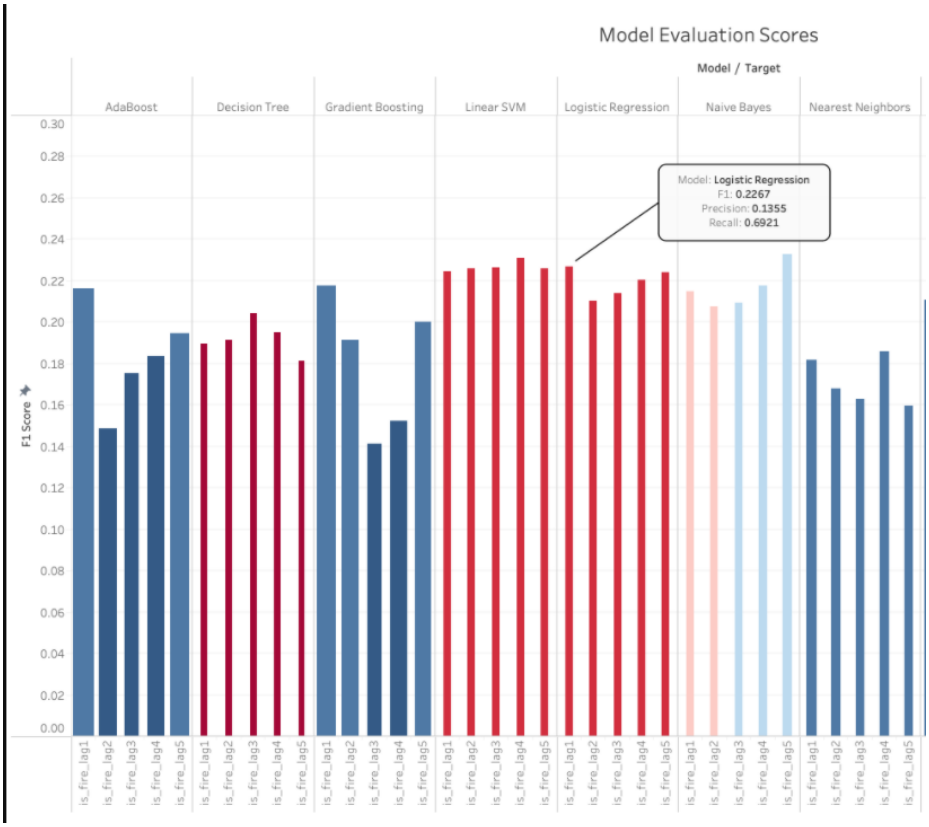

Fig. 7. Model Evaluation Scores

In addition, the feature importance shown in figure 8highlights the engineered feature of mgrs_fires_past3yrs as one of the most important features for the prediction.

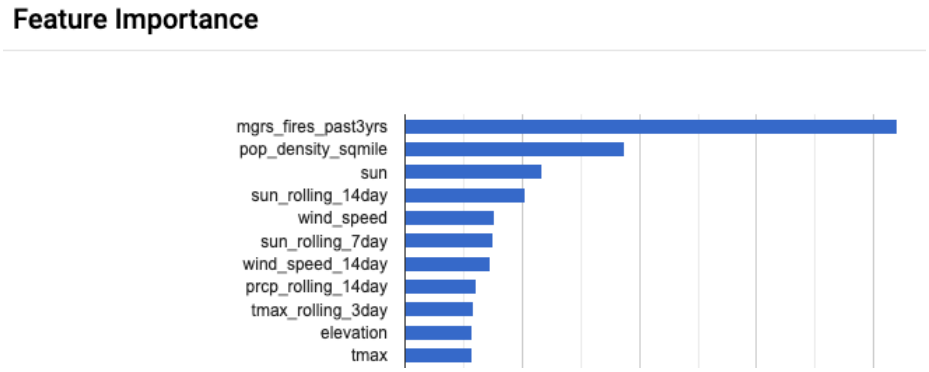

Fig. 8. Feature importance

This model can now predict the probability of fire greater than one-quarter of an acre occurring in each region within the next one to five days.

4 Results

Results were analyzed by taking weather data from the date ranges of January 1st, through June 18th, 2021, and using this data to analyze the predictions by region and time. Since real-time data on fire occurrences are not readily available, the first step in the analysis is to look at the most recent predictions by MGRS location to inspect their accuracy visually. Fires should be predicted in areas that are more prone to fires than in the past. Fires should not be predicted in areas of high urban density or coastal areas that historically have had few fire incidents.

Figure 9 below shows the areas where fires are predicted to occur within the next five days. Since June is peak fire season in southern California, many areas appear as large red circles. The areas with the highest probabilities generally lie along mountain ridges surrounding the most densely populated coastal regions. Other sites, such as the area north-south between San Clemente and Oceanside, are located in the Marine Corps Base Camp Pendleton. This area is known for its high frequency of fires since it is largely undeveloped and uninhabited. The area along the northern coastline, roughly north of Malibu, is also an area known for fire outbreaks due to the geographic proximity to mountains.

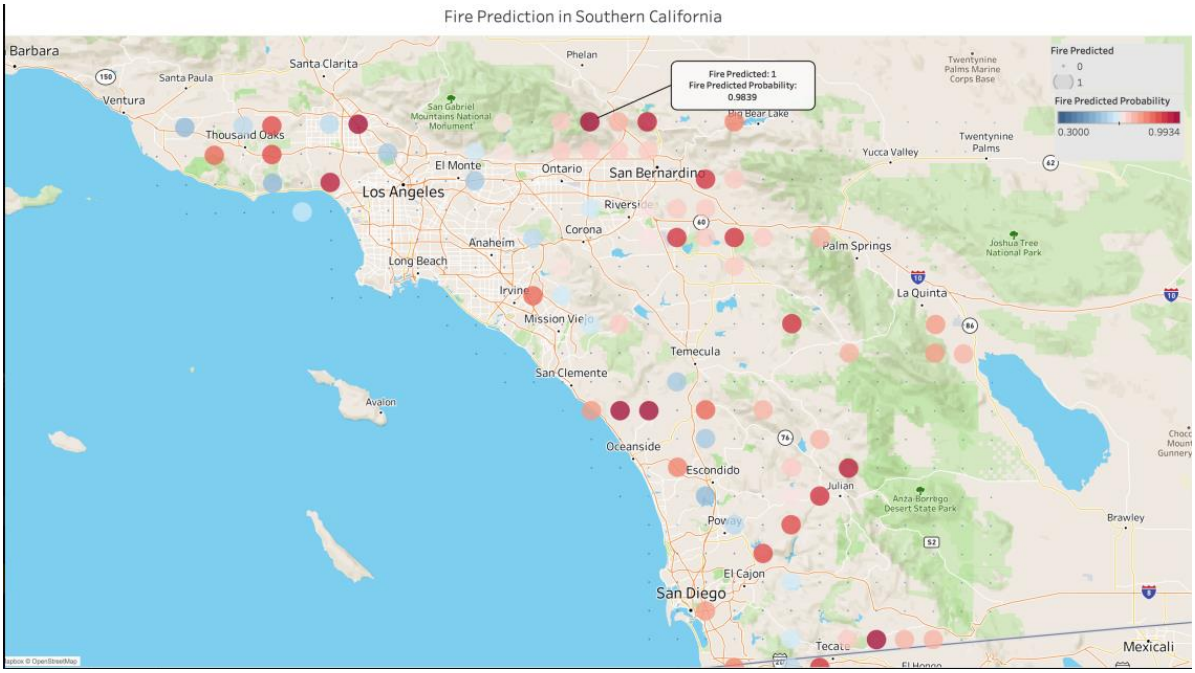

Fig. 9. Model Results – 6/18/2021

Next, the regions' highest and lowest probabilities are located. Figure 10 shows the minimum and maximum probability locations as of June 18th, 2021.

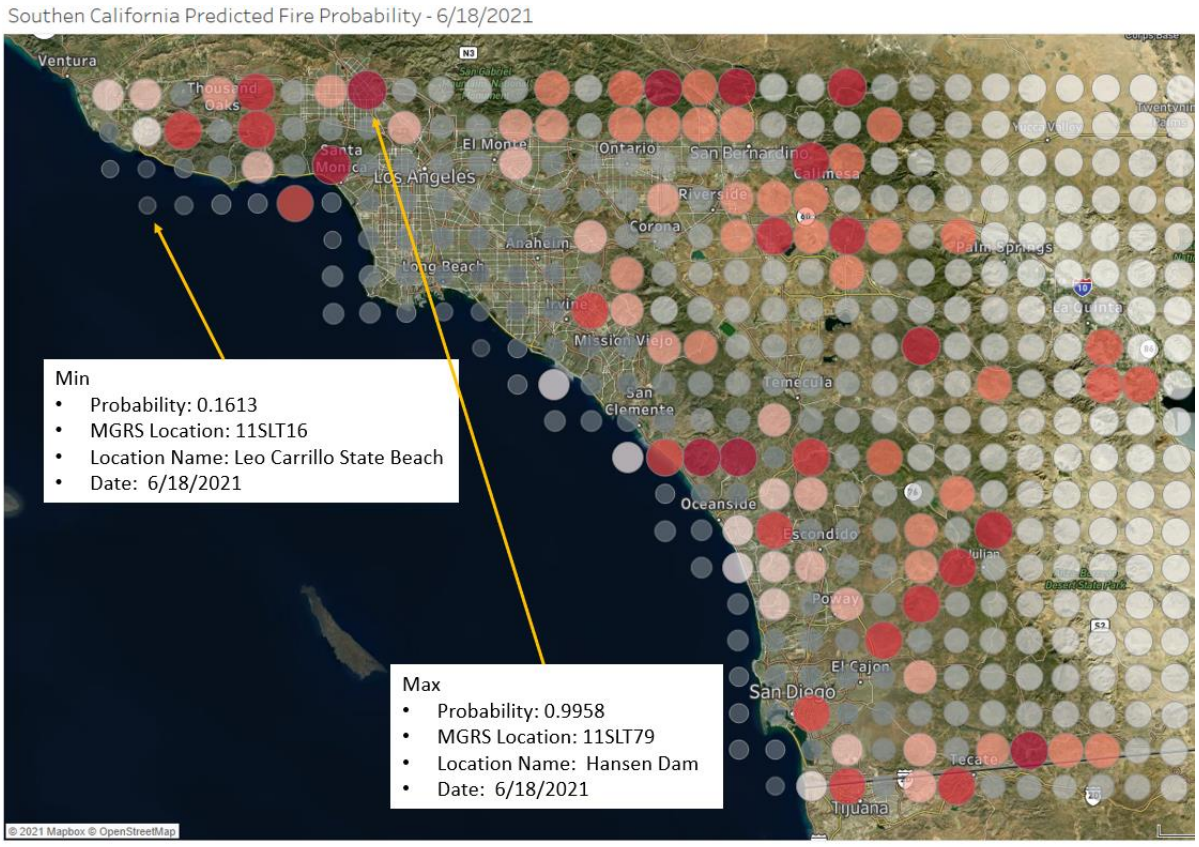

Fig. 10. Highest and Lowest Risk Areas – 6/18/2021

A closer inspection of the Hansen Dam area shows a densely populated region nestled tightly beside a large mountain range. Kagel Canyon, which is just north of Hansen Dam and within the exact MGRS grid location, was the originating location of the Creek Fire in 2017 that burned approximately 15,323 acres and 60 homes. Figures 11 and 12 below show the detailed map of the Hansen Dam area and the coverage of the 2017 Creek Fire.

Fig. 11. Hansen Dam Area Detail Map

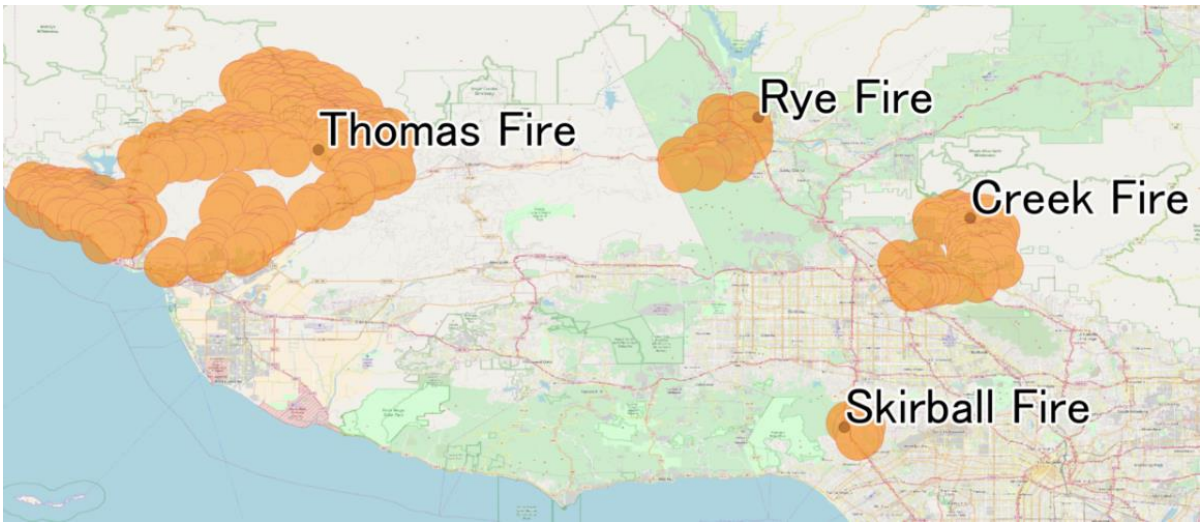

Fig. 12. Creek Fire Total Area Burned December 2017

The area with the lowest fire probability is within the heart of coastal Malibu which lies just west of Leo Carrillo State Beach. Since only a small portion of the total 10km MGRS location covers land, it makes sense that the area would have a lower overall probability of fire. Although the area just north has a history of large fires, the relative proximity to the coast and beach, along with wealthy residents and a nearby fire department, may contribute to better land management in the immediate surrounding areas, which has led to fewer outbreaks of large fires historically.

Fig. 13. Leo Carrillo State Beach

The following parameter to check in the model results is how well the predicted probabilities adjust for the year. The plots below in Figure 14 show the predicted probability (red lines) by date for 2021 through June 18th, 2021 in the Hansen Dam MGRS location versus hours of sunlight per day and average temperature (yellow and orange lines, respectively).

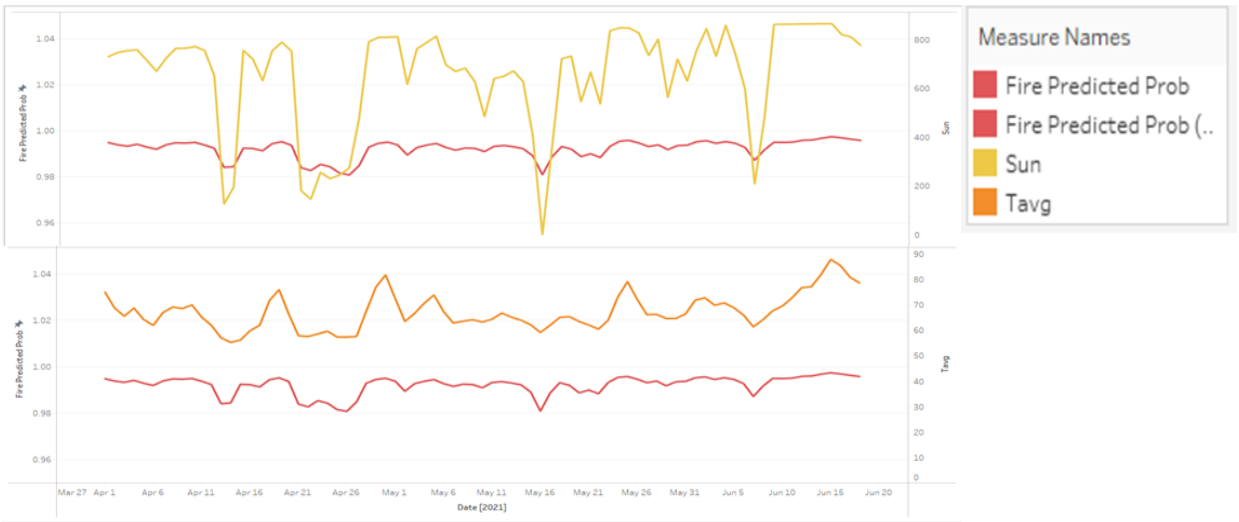

Fig. 14. Hansen Dam Predicted Fire Probability by Date - 2021

Figure 14 shows how the probabilities fluctuate in direct relation to fluctuations in sun and temperature. However, the expected trend of lower possibilities in the cooler spring months versus higher possibilities in the mid-to-late spring does not hold. In fact, the elevated probability of fire in this region stays relatively constant at around 98% regardless of the time dimension.

As the last step, the same time analysis is performed for the low probability area of Leo Carrillo State Beach.

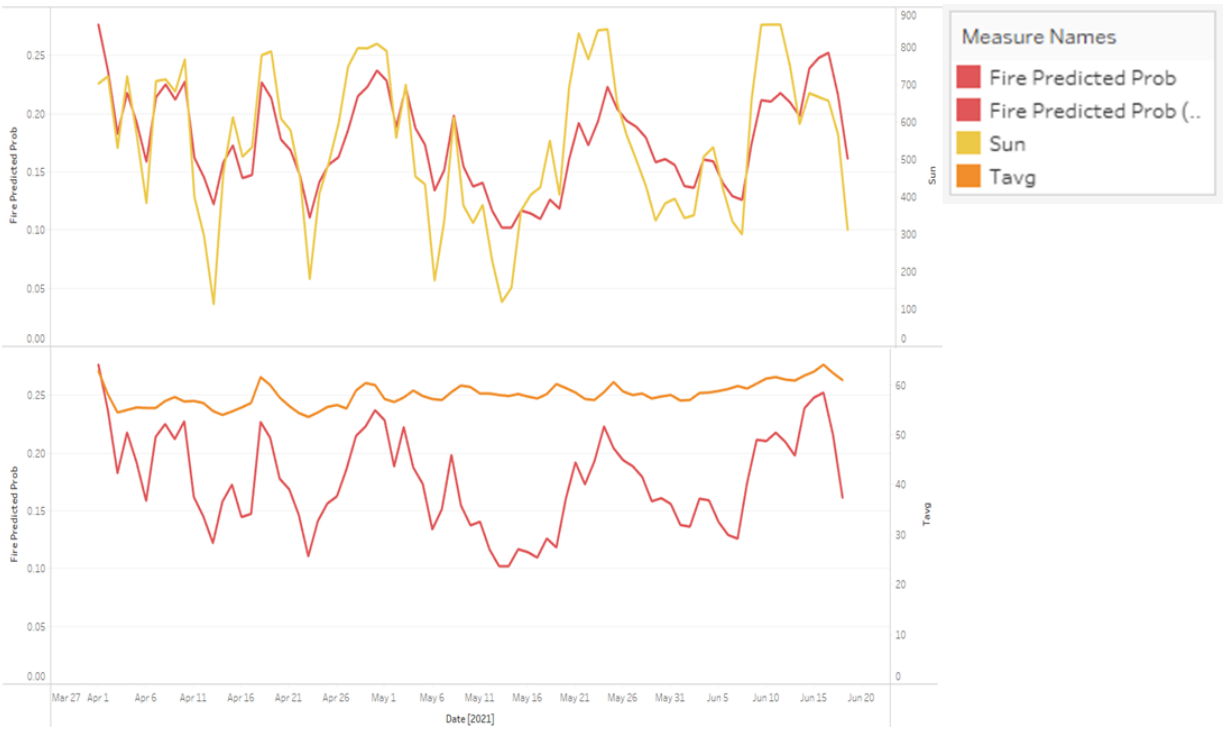

Fig. 15. Leo Carrillo State Beach Predicted Fire Probability by Date - 2021

Figure 15 above shows similar results for the highest probability area in Figure 14, where fire probabilities remain relatively constant throughout the year.

In summary, the results appear to correctly identify areas that have historically experienced high frequencies of fires. However, adding the dimension of time as a model parameter input seems necessary to better predict fires as they occur throughout the year.

5 Discussion

The interesting part about this study is the impact that it can have on saving lives and property. The results of the models can be used by government and fire officials to protect areas of concern by taking proactive measures based on the prediction of a fire occurring. The proactive measures could be clearing brush and debris away from high-value assets and taking precautionary measures to evacuate people from the area. While this study focused on the Southern California region, the model could be adapted to handle any area. As seen in Figure 16, this information can help save lives based on the population density and the volume of fires.

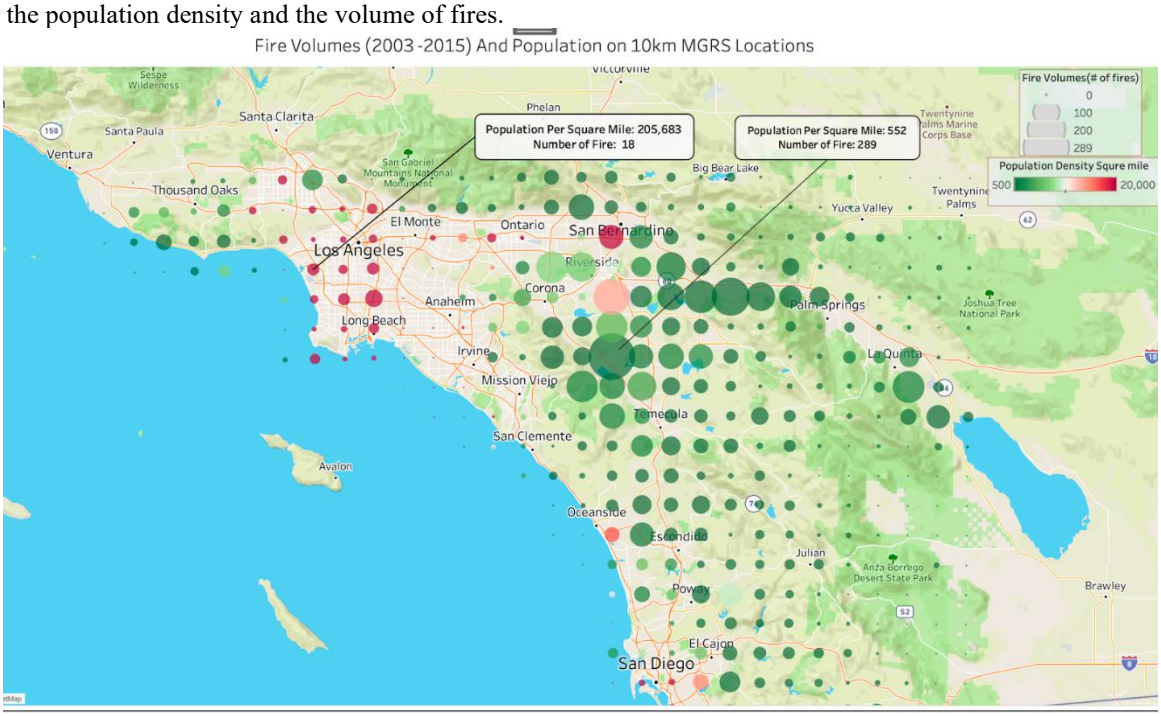

Fig. 16. Population density and fire volume for Southern California

Getting a prediction that a fire is imminent could prevent loss of life and catastrophic damage.

This study's concern from an ethical perspective is arsonists using the predictability of a fire occurring and prematurely starting a fire. If an arsonist knows that a fire is predicted for an area, this could help motivate their actions. Another concern and ethical implication surrounding the predictability of fires would be around insurance fraud. Suppose a person knew ahead of time that a fire was imminent. In that case, they could use this information to make their property more susceptible to a fire or even start the fire themselves to collect insurance money fraudulently. Because of this, the model's availability should not be made public but provided under close supervision of government officials and used for official business only.

6 Conclusion

The initial goal for this study was to provide the ability to characterize and classify areas with risk levels of forest fires to limit the loss of life, property damage and improve quality of life by allowing emergency responders to have a proactive response rather than reactive response to wildfires. The results indicate that this is possible and could be used by government and fire officials and incorporated into their process for fire prevention. An area that would require more attention is the value of the model output compared to what's currently available. There are several methods in use today for detecting fires. While the process outlined in this study isn't meant to deprecate anything already in place, it could be used to augment or enhance the current flow of information. Future work in this area could expand on the geographic regions. While the focus was on Southern California, having the predictability for other sites would also be beneficial. If another researcher were to pick this study up and take it further, the recommendation would be to focus on more feature engineering. The challenge encountered with this study is the availability of good, consistent fire data, hence the focus on a smaller area. However, with more time, better data could be curated with a data pipeline to ascertain, clean, and have the data readily available for a higher-performing model. Tighter integration with existing methods for detecting fires and this model technique for assigning risk levels and predicting fires can help save lives.

References

1. Ghorbanzadeh, Omid; Kamran, Khalil Valizadeh; Blaschke, Thomas; Aryal, Jagannath; Naboureh, Amin; Enlai, Jamshid; Bian, Jinhu; (2019). Spatial Prediction of Wildfire Susceptibility Using Field Survey GPS Data and Machine Learning Approaches. MDPI Fire, 2(3), 43 (July 2019). https://doi.org/10.3390/fire2030043.
2. Giglio, Louis; Loboda, Tatiana; Roy, David P.; Quayle, Brad; Justice, Christopher O.; (2009) An active fire based burned area mapping algorithm for the MODIS sensor. Remote Sensing of Environment, 113(2), 408-420. https://doi.org/10.1016/j.rse.2008.10.006.
3. Toulouse, Tom; Rossi, Lucile; Campana, Antoine; Celik, Turgay; Akhloufi, Moulay A. (2017). Computer vision for wildfire research: An evolving image dataset for processing and analysis. Fire Safety Journal, 92, 188-194. https://doi.org/10.1016/j.firesaf.2017.06.012.
4. Sayad, Younes Oulad; Mousannif, Hajar; Moatassime, Hassan Al (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. Fire Safety Journal, 104, 130-146.
5. Wirth, T., & Pyke, D. (2007). Monitoring post-fire vegetation rehabilitation projects: a common approach for non-forested ecosystems / by Troy A. Wirth and David A. Pyke U.S. Geological Survey.
6. Belcher, C., & Abiven, S. (2014). Fire phenomena and the earth system: an interdisciplinary guide to fire science / edited by Claire M. Belcher; contributors, Samuel Abiven [andtwenty-four others]. Wiley-Blackwell.
7. Sen, Z. (2015). Applied drought modeling, prediction, and mitigation. ProQuest eBook Central https://ebookcentral-proquest-com.proxy.libraries.smu.edu.
8. Andrews, P., Finney, M., & Fischetti, M. (2007). Predictingwildfiress. Scientific American, 297(2), 46-55. Retrieved from https://www.jstor.org/stable/26069414?seq=1.

9. B. J. Stocks, J. A. Mason, J. B. Todd, E. M. Bosch, B. M. Wotton, B. D. Amiro, . . . W. R. Skinner. (2002). Large forest fires in Canada, 1959–1997. Journal of Geophysical Research-Atmospheres, 107(D1), 8149-FFR 5. doi:10.1029/2001JD000484.
10. Cortez, P., & Morais, A. J. R. (2007). A data mining approach to predict forest fires using meteorological data Associação Portuguesa para a Inteligência Artificial (APPIA). Retrieved from http://hdl.handle.net/1822/8039.
11. Haiganoush K. Preisler, & Anthony L. Westerling. (2007). Statistical model for forecasting monthly large wildfire events in western united states. Journal of Applied Meteorology and Climatology, 46(7), 1020-1030. doi:10.1175/JAM2513.1.
12. Reid, C. E., Jerrett, M., Petersen, M. L., Pfister, G. G., Morefield, P. E., Tager, I. B., . . . Balmes, J. R. (2015). Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning, American Chemical Society (ACS). doi:10.1021/es505846r.
13. Stephen J. Pyne. Passing the Torch: Why the eons-old truce between humans and fire has burst into an age of megafires, and what can be done about it. The American Scholar Vol. 77, No. 2 (SPRING 2008), pp. 22-32 (11 pages). https://www.jstor.org/stable/41222727?mag=the-age-of-megafires&seq=3#metadata_info_tab_contents.
14. Janet Franklin, Alexandra D. Syphard, Hong S. He and David J. Mladenoff. Altered Fire Regimes Affect Landscape Patterns of Plant Succession in the Foothills and Mountains of Southern California. Ecosystems Vol. 8, No. 8 (Dec. 2005), pp.885-898(14pages). https://www.jstor.org/stable/25053885?mag=fire-season-getting-longer-longer&seq=1#metadata_info_tab_contents.
15. Max A. Moritz, Jon E. Keeley, Edward A. Johnson, and Andrew A. Schaffner. Testing a Basic Assumption of Shrubland Fire Management: How Important Is Fuel Age? Frontiers in Ecology and the Environment Vol. 2, No. 2 (Mar.,2004), pp.67-72(6pages). https://www.jstor.org/stable/3868212?mag=fire-season-getting-longer-longer&seq=1#metadata_info_tab_contents.
16. Marc Macias Fauria and E. A. Johnson. Climate and Wildfires in the North American Boreal Forest. Philosophical Transactions: Biological Sciences Vol. 363, No. 1501, The Boreal Forest, and Global Change (Jul. 12, 2008), pp. 2317-2329(13pages). https://www.jstor.org/stable/20208640?mag=how-forest-fires-work-in-finland&seq=1#metadata_info_tab_contents.
17. Steblein, P.F., Loehman, R.A., Miller, M.P., Holomuzki, J.R., Soileau, S.C., Brooks, M.L., Drane-Maury, M., Hamilton, H.M., Kean, J.W., Keeley, J.E., Mason, R.R., Jr., McKerrow, A., Meldrum, J.R., Molder, E.B., Murphy, S.F., Peterson, B., Plumlee, G.S., Shinneman, D.J., van Mantgem, P.J., and York, A., 2021, U.S. Geological Survey wildland fire science strategic plan, 2021–26: U.S. Geological Survey Circular 1471, 30 p., https://doi.org/10.3133/cir1471.
18. Steblein, P.F., and Miller, M.P., 2018, Characterizing 12 years of wildland fire science at the U.S. Geological Survey—Wildland fire science publications, 2006–17: U.S. Geological Survey Open-File Report 2019–1002, 67 p., https://doi.org/10.3133/ofr20191002.
19. Steblein, P.F., Miller, M.P., and Soileau, S.C., 2019, Wildland fire science at the U.S. Geological Survey—Supporting wildland fire and land management across the United States (ver. 2.0, February 2021): U.S. Geological Survey General Information Product 190 [postcard], 2 p., https://doi.org/10.3133/gip190.
20. Steblein, P.F., Miller, M.P., and Soileau, S.C., 2019, Wildland fire science---Supporting wildland fire and land management: U.S. Geological Survey Fact Sheet 2019–3025, 2 p., https://doi.org/10.3133/fs20193025.
21. Michael M. Arms and John D. Van Zante. Wildfire Evacuation: Outrunning the Witch's Curse-One Animal Center's Experience. ILAR journal, 2010, Vol.51 (2), p.158-163. https://doi-org.proxy.libraries.smu.edu/10.1093/ilar.51.2.158.

22. M.D.Flannigan, B.J. Stocks, B.M., 2000. Wotton. Climate change and forest fires. The science of the total environment, 2000, Vol.262 (3), p.221-229. https://doi.org/10.1016/S0048-9697(00)00524-6.
23. Cheuklap Au Yeung & Rongrong Li. Comparison of vegetation regeneration after wildfire between Mediterranean and tundra ecosystems by using Landsat images, Annals of GIS Volume 24, 2018, pp. 99-112. https://www.tandfonline.com/doi/full/10.1080/19475683.2018.1424740
24. Heris, M.P., Foks, N., Bagstad, K., and Troy, A., 2020, A national dataset of rasterized building footprints for the U.S.: U.S. Geological Survey data release, https://doi.org/10.5066/P9J2Y1WG.
25. Inter-University Consortium for Political and Social Research (ICPSR) Duranton, Gilles, Puga, Diego. (2020, May 5). Data and code for: The economics of urban density [Dataset]. American Economic Association. https://www.openicpsr.org/openicpsr/project/119268/version/V1/view?path=/openicpsr/119268/fcr:versions/V1&type=project.
26. VanSomeren, L., Nelson, H., Burn, L. R., & Joseph, K. (2020, February 7). The environmental impact of forest fires. Untamed Science. https://untamedscience.com/blog/the-environmental-impact-of-forest-fires/.